Piotr Domagała

SEGMENTALNE ROZPOZNAWANIE FONOKODU DO STEROWANIA MASZYNĄ ROBOCZĄ

11/1987

WARSZAWA 1987

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 25 listopada 1986 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN

Nakład 140 egz. Ark.wyd. 1 Ark.druk. 1,5
Oddano do drukarni w marcu 1987 r.
Nr zamówienia 156/87.

Warszawska Drukarnia Naukowa, Warszawa,
ul. Śniadeckich 8

Piotr Domagała
Pracownia Fonetyki Akustycznej
IPPT PAN

SEGMENTALNE ROZPOZNAWANIE FONOKODU DO STEROWANIA MASZYNĄ ROBOCZĄ[1]

Streszczenie.

Przedstawiono i uzasadniono dobór fonemów bazowych (i, a, u, s, p) będących podstawą do konstrukcji fonokodu oraz sformułowano zasady jego budowy. Przeprowadzono eksperymenty nad rozpoznawaniem 18 haseł wypowiedzianych przez 8 głosów. Uzyskano 94 % poprawność rozpoznawania niezależnie od głosu.

1. Wstęp.

Poprawna praca maszyny zależy od niezawodnego działania jej podzespołów, stanu obiektu na który oddziaływuje oraz od poprawnego sterowania, które zapewnia odpowiednio dobrany algorytm i wartości sygnału sterującego. Jeżeli sygnał sterujący jest elektryczny, pneumatyczny, hydrauliczny lub mechaniczny, to utrzymanie jego parametrów we właściwym zakresie uwarunkowane jest poprawną pracą nadajnika informacji oraz zabezpieczeniem toru do odbiornika przed zakłóceniami i zniekształceniami. Sygnały tego typu posiadają najczęściej standardowe wartości. Sygnałem sterującym o szczególnej specyfice jest mowa. Obok obrazu, pisma i rzeźby służy ona człowiekowi do przekazywania swoich myśli. Odpowiednio dobrane i wypowiedziane jej elementy stanowią ogromne bogactwo środków zapewniających precyzyjne zakodowanie nawet najbardziej subtelnych

[1] Praca wykonana w ramach CPBP 02.13

odczuć. W komunikacji człowiek-człowiek informacje zawarte w naturalnej wypowiedzi mogą zostać poprawnie odebrane pomimo dużego poziomu zakłóceń, zniekształceń i niejednoznaczności. Człowiek, aktywnie uczestnicząc w procesie odbioru wykorzystuje swoją wiedzę a pełniący funkcję informacyjnego sprzężenia zwrotnego dialog podwyższa niezawodność kanału łączności. Dorosła osoba wykorzystuje przeciętnie 12000 słów. Dziesięciotomowy Słownik Języka Polskiego pod red. W. Doroszewskiego zawiera około 125000 wyrazów. Liczby te są nieznaczne w porównaniu z teoretycznymi możliwościami jakie daje 40 fonemowy język polski. Prosty rachunek dowodzi, że liczba słów zbudowanych co najwyżej z k fonemów wynosi $\sum_{i=1}^{k} 40^i$ (dla k=10 liczba ta przekracza wartość 16×10^{18}). Przedstawione obliczenie ma znaczenie tylko poglądowe, gdyż dopuszcza występowanie takich "słów", których poprawna wymowa jest bardzo trudna lub niemożliwa z uwagi na trudności artykulacyjne np. sdtętopz.

W wieloetapowym procesie automatycznego rozpoznawania mowy (ARM) przez maszynę, która nie posiada inteligencji na poziomie zbliżonym do ludzkiej, szczególnie ważny jest etap akustyczny. Nie jest rzeczą wykluczoną, że dotychczasowy brak pełnej akustycznej teorii fonemu oraz tylko niepełne modele inteligencji uniemożliwiają swobodną rozmowę człowieka z maszyną /Jassem 1985/. Pomimo tych trudności realizowane są układy, które umożliwiają komunikację głosową z maszyną. Działanie większości komercyjnych układów ARM polega na rozpoznawaniu izolowanych wyrazów (komend), które mogą być analogami przycisków sterowniczych. Liczba identyfikowanych wyrazów (od kilku do kilku tysięcy) wpływa bezpośrednio na koszty urządzenia, czas reakcji i stopień pewności działania.

Niezależnie od typu kanału przekazywania informacji liczba komend, którymi człowiek steruje maszyną jest skończona. Nawet sterowanie analogowe np. kierownicą samochodu można poprzez kwantyzację zamienić na skończoną liczbę dopuszczalnych wartości. Każda komenda musi zawierać w sobie element wyboru (adres podzespołu) oraz element działania. Elementy te mogą być oddzielnymi decyzjami operatora np. "zawór nr 3 - ot-

wórz" lub połączone w jednej postaci. Wśród rozkazów wyróżnić należy grupę o szczególnie ważnych priorytetach - polecenia o charakterze alarmowym o podwyższonej pewności zadziałania. Jeżeli sygnał mowy jest sygnałem sterującym, to poszczególnym komendom umożliwiającym sterowanie maszyną odpowiadają poszczególne wyrazy, które tworzą pewien ograniczony słownik. Układ ARM powinien w sposób niezawodny dekodować wydawane rozkazy, co oznacza poprawną ich detekcję przy eliminacji zróżnicowań wynikających z przyczyn losowych lub patologicznych. Odrębnym problemem jest występowanie cech osobniczych w sygnale mowy. W zależności od celu może istnieć potrzeba ich pomijania (wówczas sterowanie maszyną jest ogólnie dostępne) lub wyodrębnienia, uniemożliwiając dostęp do maszyny osobom postronnym. Niniejsza praca pomija zagadnienie ekstrakcji cech osobniczych.

Operator, dążąc do uzyskania pożądanej reakcji układu technicznego, może sformułować swój zamiar kilkoma sposobami, np. "stop", "stój", "zatrzymaj","stań" itp. Przy aktualnych realiach technicznych nieekonomiczne byłoby, ażeby tej samej czynności przyporządkować różne polecenia. W takiej sytuacji procedury detekcyjne i dyskryminacyjne ulegają skomplikowaniu a prawdopodobieństwo nieprawidłowej interpretacji rozkazu zwiększa się.

Odpowiedni dobór rozpoznawanych słów jest ważnym czynnikiem wpływającym na poziom sprawności układu ARM. Występowanie wyrazów podobnych "sto", "stop" może być przyczyną błędnej identyfikacji. Procedura rozpoznawania ulega redukcji przy wzroście stopnia kontrastywności rozpoznawanych haseł. Dreyfus-Graf /1974/ zaproponował konstrukcję sztucznej mowy, zwanej fonokodem, która znacznie lepiej jest rozpoznawana przez maszynę niż naturalna. System fonetyczny języka naturalnego został ograniczony do podzbioru zróżnicowanych fonemów (s,o,t,i,n,a,k, e,m,u,z). Istnieje wówczas możliwość przyporządkowania każdemu naturalnemu fonemowi ciągu fonemów należących do tego podzbioru. Umożliwia to przekazanie każdej treści przy kilkukrotnym

wydłużeniu czasu wypowiedzi (Koster, Dreyfus-Graf/1976/).
W niniejszej pracy idea Dreyfusa-Grafa została wykorzystana do budowania łatwo rozpoznawalnych haseł.

2. Zasady konstrukcji fonokodu.

Fonokod, rozumiany w tej pracy jako wartość funkcji przyporządkującej niektórym pojęciom lub wyrazom sztuczne konstrukcje leksykalne, powinien spełniać poniższe wymagania.
A. Wszystkie utworzone sztuczne wyrazy zbudowane są z ograniczonego zbioru fonemów. Warunek ten upraszcza procedurę rozpoznawania i skraca czas jej wykonywania.
B. Fonemy, w oparciu o które tworzone są hasła, są maksymalnie zróżnicowane pod względem akustycznym i fonologicznym.
C. Artykulacja logatomów nie może stwarzać żadnych trudności.
D. Hasła, które znajdują zastosowanie w układach praktycznych, powinny być dostatecznie zróżnicowane. Pamięciowe opanowanie listy kilkunastu (kilkudziesięciu) poleceń nie będzie stanowić problemu.

3. Wybór podzbioru fonemów.

3.1 Samogłoski.

W języku polskim występuje 6 podstawowych samogłosek : i, ɨ, e, a, o, u. Zarówno badania percepcyjne, jak i analiza struktur widmowych wskazują, że zasadnicze cechy dyskryminacyjne są zlokalizowane w obrębie 2 pierwszych formantów. Rozkład samogłosek na czworoboku artykulacyjnym przedstawia rys.1. Najbardziej skrajne położenia zajmują "i", "a", "u". Rys.2 (wg Jassema /1984/) przedstawia położenie średnich częstotliwości dwóch pierwszych formantów dla 4 głosów . Również z tego rysunku wynika, że najbardziej oddalone od siebie są te same samogłoski. Fonemy te zostały wybrane do konstrukcji fonokodu.

3.2 Spółgłoski.

Do grupy spółgłosek posiadających segment zwarcia należą : t͡s, d͡z, t͡ɕ, d͡ʑ, t͡ʃ, d͡ʒ, p, b, t, g, k, ɟ, c. Założono, że właśnie ten segment będzie decydował o detekcji fonemu. Ponieważ spółgłoski te są polisegmentalne, odrzucono te, w których różnica między segmentem zwarcia a innymi segmentami jest stosunkowo mała, jak jest w przypadku głosek zwarto-trących. Detekcję zwarcia naj-

prościej zrealizować poprzez wykrycie zaniku energii w wyrazie, co następuje w głoskach bezdźwięcznych p,t,k,c . Spośród nich głoska p posiada najkrótszy segment szumów poplozyjnych por. (Jassem /1973/ str.240). Z grupy spółgłosek trących s,ʃ,ɕ, f,x, z,ʒ,ʑ, v wybrano "s". W dźwięcznych fonemach składowe tonu podstawowego mogą zachodzić na obszar występowania pierwszego formantu samogłosek.

Odrzucono grupę głosek nosowych. Ich rozróżnienie i automatyczna identyfikacja są problematyczne przy dostępnym systemie pomiarów i analizy . Obrazy widmowe głosek "j","w","l" są podobne do analogicznych obrazów samogłosek i,u,a. Drżąca realizacja głoski r utrudnia zarówno segmentację, jak i jednoznaczną interpretację spektrogramu, natomiast segment uzyskiwany w wyniku realizacji uderzeniowej np. kark jest zbyt krótki. Reasumując : ze zbioru fonemów występujących w języku polskim wybrano 5 głosek do konstrukcji fonokodu - "i","a","u","p","s".

4. Konstrukcja fonokodu.

Ograniczenie zbioru fonemów z 39 do 5 jest ważnym, ale nie jedynym możliwym czynnikiem, który ma wpływ na jakość procesu ARM. Dokonano następujących założeń ograniczających łączenie wybranych fonemów :

- głoski występujące w fonokodzie muszą spełniać zasady fonotaktyki języka polskiego,
- fonem p nie może być pierwszym ani ostatnim w logatomie,
- samogłoski nie mogą występować w bezpośrednim sąsiedztwie.

Metodę konstrukcji fonokodu można przedstawić graficznie rys.3 Strzałki wskazują możliwość doboru następnego fonemu, np. jeśli ostatnio dobrany fonem był a, to następny może być s lub p. W Tablicy 1 przedstawiono liczbę możliwych do utworzenia logatomów w zależności od długości fonokodu. Fonokody zestawiono w Załączniku.

5. Opis metody segmentalnego rozpoznawania fonokodu.

5.1 Metoda automatycznej segmentacji mowy ASM .

Do realizacji metody wykorzystano układ analogowo-cyfrowy, w skład którego wchodzą : 60-kanałowy analizator widma,

interface łączący analogowe źródło sygnału z minikomputerem, minikomputer MERA 303. Analogowy alalizator widma posiada 43 pasma analizy o stałej szerokości wynoszącej 80 Hz pokrywające zakres częstotliwości od 120 Hz do 3560 Hz oraz 17 pasm o szerokości zależnej od częstotliwości środkowej w sposób liniowy i pokrywających zakres od 3560 Hz do 7000 Hz. Wyjścia poszczególnych kanałów są cyklicznie załączane na wspólny tor wyjściowy. Otrzymany spektrogram cyfrowy sygnału mowy, po uśrednieniu zostaje zapamiętany w pamięci minikomputera jako tablica o współrzędnych czasu i częstotliwości (Domagała /1984/). Dla każdej pary kolejnych widm k-1 i k utworzono ciąg N-wyrazowy N-liczba kanałów o elementach $r_{ik} = a_{ik-1} - a_{ik}$, gdzie i=1,...,N oznacza numer pasma, k kolejny kwant czasu. Ciąg (r_{ik}) podzielono na c(k) ciągów składowych stosując kryterium zgodności znaku i dostatecznie dużej wartości bezwzględnej, to jest kierunku i prędkości zmian poziomu. Jako wartość progową prędkości zmian poziomów sygnału w poszczególnych kanałach analizatora widmowego przyjęto ok. 30 dB/23 ms 23 ms jest to odległość czasowa między kolejnymi widmami . Wartość ta była stała dla wszystkich głosów. Jako z(k) oznaczono znak wyrazów ostatniego ciągu składowego, przyjmując 0 dla wartości dodatnich i 1 dla ujemnych. Przyjęto, że granica między segmentami zostanie położona w następujących przypadkach :

$$c(k) = 1 \wedge \overline{c(k+1)} \neq 1 \vee \overline{c(k+1)} = 1 \wedge z(k) \neq z(k+1) \quad (1)$$

$$c(k) = c(k+i) \neq 1 \wedge z(k+i) = z(k) \wedge [\overline{c(k+n)} \neq 1 \vee \overline{c(k+n)} = 1 \wedge z(k) \neq z(k+n)]] \quad (2)$$

gdzie i=1,2,...,n-1

$$c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) \neq 1 \wedge c(k+1) \neq 2 \quad (3)$$

$$c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) \neq 1 \wedge c(k+i) = 2 \wedge z(k) = z(k+i) \wedge z(k) = z(k+n) \wedge c(k+n) = 2 \quad \text{gdzie } i=0,1,...,n-1 \text{ oraz } n \geq 1 \quad (4)$$

$$c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) \neq 1 \wedge c(k+i) = 2 \wedge z(k) = z(k+i) \wedge c(k+n) \neq 1 \wedge c(k+n) \neq 2 \quad \text{gdzie } i=0,1,...,n-1 \text{ oraz } n \geq 2 \quad (5)$$

Powyższe zależności logicznie implementowano w systemie MERA 303.

5.2 Ekstrakcja segmentów.

W wyniku przedstawionych powyżej procedur obraz widmowy wypowiedzi podzielony został na segmenty, które w przybliżeniu

odpowiadają fonemom. Dla każdego segmentu obliczono jego średnie widmo SW wg poniższej zależności :

$$SW_K = \sum_{i=n_1}^{n_2-1} W_i(k) \frac{1}{n_2 - n_1 - 1} \qquad (6)$$

gdzie k oznacza nr kanału analizatora widma (k=1...60), $W_i(k)$ oznacza wartość widma w k-tym kanale i-tej chwili czasu oraz n_1 i n_2 są granicami segmentu.

Ponadto nałożono ograniczenia na iloczasy segmentów, które polegają na pomijaniu granic położonych zbyt blisko siebie segmenty o czasie trwania mniejszym niż 60 ms oraz w przypadku zbyt dużej rozpiętości czasowej segmentu większej niż 230 ms na wprowadzeniu sztucznej granicy w połowie tego obszaru. Liczne obserwacje potwierdziły słuszność takiej modyfikacji. Znaczna część granic nadmiarowych zostaje w ten sposób usunięta, natomiast tam gdzie występowały braki segmentacji, wprowadzona sztuczna granica na ogół trafnie je uzupełniała. Rys.4 przedstawia średnie widma segmentów głosek "i", "a", "s", "u" dla 4 głosów i trzykrotnej realizacji każdej głoski.

W kolejnym kroku dokonano normalizacji segmentów SW, która polegała na takim przekształceniu SW , ażeby uzyskać stałą wartość średnią :

$$SW1_K = \frac{const}{\frac{1}{N}\sum_K SW_K} \cdot SW_K \qquad (7)$$

gdzie SW1 k oznacza znormalizowaną wartość średniego widma segmentu w k-tym kanale, N - liczbę kanałów, const (tu const=24) stałą będącą wspólną średnią.

Rys.5 przedstawia średnie, unormowane widma głosek przedstawionych na rys.4.

W dalszym etapie zredukowano liczbę danych reprezentujących poszczególne fonemy z 60 do 12 poprzez uśrednienie arytmetyczne wartości występujących w każdych kolejnych 5 kanałach.

Na rys.6 wykreślono uśrednione po wszystkich realizacjach, u-

normowane i zredukowane reprezentacje głosek "i", "a", "s", "u", które przyjęto jako zbiór wektorów wzorcowych w celu porównania z segmentami różnych wypowiedzi w procesie rozpoznawania fonokodu. Segmenty te, ekstrahowano w analogiczny sposób, tzn. automatycznie uzyskiwane segmenty uśredniano, normalizowano i redukowano. Fragmenty sygnału mowy o znacznie obniżonym poziomie energetycznym przed normalizacją interpretowano jako wystąpienie głoski "p". Mając na uwadze przyszłą hardware'ową realizację systemu, wszystkie obliczenia przeprowadzono na liczbach całkowitych.

5.3 Rozpoznawanie segmentów.

Założono, że proces rozpoznawania segmentów będzie dokonywany w 3 etapach. W pierwszym etapie obliczono odległość między rozpoznawanym fragmentem fonokodu a poszczególnymi wzorcami przyjmując miarę odległości wyrażoną wzorem (8).

$$d = \sum_{i=1}^{12} |SW1_{iW} - SW1_{iR}| \qquad (8)$$

gdzie indeksy W i R dotyczą odpowiednio wzorca i segmentu rozpoznawanego.

Dla każdej części analizowanej wypowiedzi otrzymywano cztery wartości odległości od wzorców. Przyjęto punktową regułę decyzyjną, która przyporządkowuje każdemu segmentowi fonem, którego wzorzec jest najbliżej położony wg miary (8). Wynikiem pierwszego etapu jest ciąg głosek. W etapie drugim dokonywano weryfikacji syntaktycznej wyników rozpoznawania etapu poprzedniego. Z otrzymanego ciągu głosek wyeliminowane są wszystkie głoski "p" występujące na początku i końcu wypowiedzi. Diady zbudowane z tych samych fonemów zredukowano do pojedynczych głosek.

W następnym kroku przeprowadzana jest analiza diad, której celem jest detekcja połączeń z założenia uznanych za niedopuszczalne. W tym momencie następuje pierwsza sygnalizacja błędnego rozpoznania. W procesie sterowania maszyną jest to bardzo istotny moment. Operator otrzymuje informację binarną, że jego polecenie jest źle interpretowane przez maszynę. Jego kolejną

czynnością powinno być powtórzenie rozkazu.
Hasła, których formalna struktura została uznana za poprawną, poddawane są w etapie trzecim weryfikacji leksykalnej, której zadaniem jest sprawdzenie, czy odebrany wyraz należy do założonego ograniczonego zbioru rozpoznawanych haseł.

Najbardziej krytyczną jest sytuacja, w której wypowiedziany rozkaz zostaje zinterpretowany jako inny, ale też należący do słownika maszyny. Skutecznym zabiegiem eliminującym taką sytuację będzie wprowadzenie informacyjnego sprzężenia zwrotnego polegającego na przekazaniu do operatora informacji o haśle, które rozpoznała maszyna. Informacja taka może być przesłana wizualnie (napis) lub akustycznie (mowa syntetyczna). Bardziej naturalną jest druga możliwość a ponadto posiada tę zaletę, że nie absorbuje wzroku operatora dając mu możliwość swobodnej obserwacji układu maszyna-obiekt.

Przeprowadzono 2 eksperymenty nad automatycznym rozpoznawaniem fonokodu. Wykorzystano 8 głosów - 4 męskie i 4 żeńskie. Spośród nich 3 głosy brały udział w tworzeniu wzorcowych segmentów. W pierwszym eksperymencie każda z osób wypowiedziała 18 logatomów w sposób zbliżony do własnego naturalnego sposobu mówienia. Uzyskano bardzo złe wyniki rozpoznawania mniej niż 50% poprawnie rozpoznanych logatomów dla wszystkich głosów.
W doświadczeniu drugim lektorzy czytali listę logatomów powoli, w sposób zdecydowany i staranny. Zwrócono szczególną uwagę na pełną wypowiedź każdego fonemu składowego fonokodu. Wyniki rozpoznawania przedstawia tablica 2.

6. Omówienie wyników. Wnioski.

Można wymienić kilka przyczyn, które złożyły się na niepowodzenie pierwszego eksperymentu. Wprawdzie segmentacja w zdecydowanej większości przypadków była prawidłowa, to występienie braku segmentacji powodowało ekstrakcję i uśrednienie dwóch sąsiednich głosek, co uniemożliwiało ich dalsze rozpoznawanie. Zasadniczą jednak przyczyną nieprawidłowego rozpoznawania segmentów były małe wartości ich iloczasów. W większości segmentów występowały niekorzystne proporcje między stanem quasiustalonym a obszarem przejściowym, wynikające z silnego oddziaływania

uwarunkowań kontekstowych, które występują w mowie naturalnej. Niewątpliwą poprawę wyników rozpoznawania można uzyskać poprzez wprowadzenie innego sposobu obliczania widma średniego segmentu. Zamiast średniej arytmetycznej (6) należy zastosować średnią ważoną (9) celem osłabienia wpływów kontekstowych.

$$WS_k = \sum_{i=n_1}^{n_2-1} \frac{a_i \, W_i(k)}{(n_2-n_1-1) \sum_{i=n_1}^{n_2-1} a_i} \qquad (9)$$

gdzie $\max\{a_i\} = a_{\frac{n_2-n_1}{2}}$

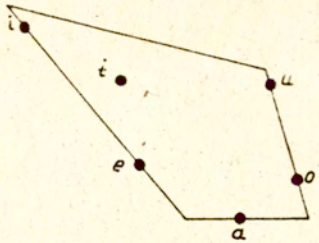
Zaproponowana powyżej modyfikacja zostanie zweryfikowana po uruchomieniu transmisji danych między systemem pomiarowym a mikrokomputerem. Nie zauważono różnic między wynikami rozpoznawania fonokodu wypowiadanego przez głosy biorące udział w konstrukcji wzorców a wynikami, które uzyskano dla innych głosów.

Prostym sposobem poprawienia relacji między czasami ustalonych części segmentów a częściami przejściowych jest spowolnienie tempa wypowiedzi. Uzyskano w ten sposób znacznie lepsze wyniki rozpoznawania - 94% poprawnych rozpoznań tablica 2 . Jednakowo dobrze rozpoznawane były logatomy wypowiadane przez głosy wzorcowe jak i pozostałe. W czterech przypadkach stan przejściowy w diadach "su" i "us" wyodrębniony został jako oddzielny segment i rozpoznany jako "i" a w jednym jako "a". Sygnalizowany był wówczas błąd (niedopuszczalna sekwencja dwóch samogłosek). W dwóch przypadkach segment składający się z plozji i szumu aspiracji został zakwalifikowany jako "s" (logatomy APU i SPU wypowiadane przez głosy RB i AM). Błędne rozpoznanie zasygnalizowano na etapie weryfikacji leksykalnej.

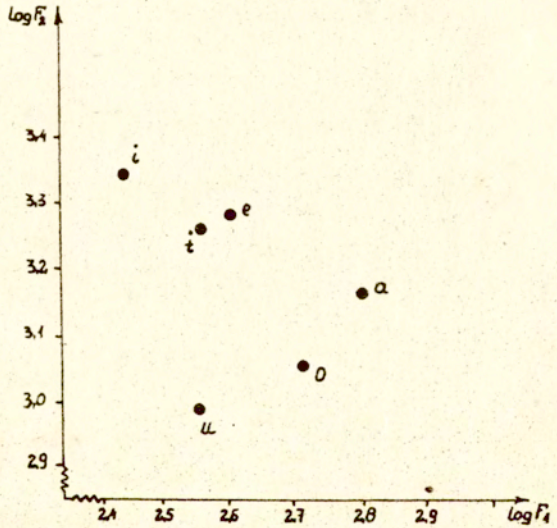
Można wyciągnąć następujące wnioski. Zastosowanie segmentalnej metody rozpoznawania fonokodu w dużym stopniu ułatwia proces rozpoznawania logatomów a w konsekwencji konstrukcję układu rozpoznawania. Układ taki będzie niewrażliwy na zróżnicowania międzyosobnicze występujące w głosie. Nadzwyczaj interesujące jest wykorzystanie techniki fonokodu w telemechanice i telemetrii. Każdemu z fonemów bazowych można przyporządkować 3 bitowy kod, który podlegałby transmisji. Z prostych obliczeń szacunkowych

wynika, że zdalne sterowanie głosem z wykorzystaniem fonokodu jest możliwe przy prędkości przesyłu wynoszącej zaledwie 10 bitów/s. Ciekawie rysują się możliwości zastosowania techniki fonokodu w telekomunikacji. Jeżeli każdemu fonemowi języka naturalnego przyporządkowany zostanie odpowiedni fonokod (zbudowany z 4-5 fonemów bazowych), to wymagana prędkość transmisji sygnału mowy wynosi około 50 bitów/s oczywiście problemem jest przekodowanie tekstu naturalnego na postać fonokodową [1]. Po stronie odbiorczej można zastosować allofoniczny syntetyzator mowy. Transmitowana w ten sposób mowa byłaby utajniona, co może w niektórych zastosowaniach mieć szczególne znaczenie.

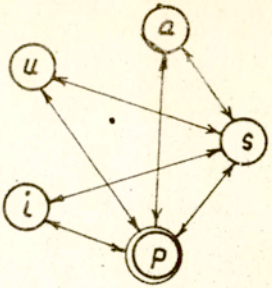
[1] Np. jeżeli fonemom języka naturalnego z, a, w, o̰, t͡ʃ przyporządkowane zostaną fonokody IPIS, APUS, ASPA, SPIS, UPIS to słowo "załącz" będzie miało postać IPIS APUS ASPA SPIS UPIS. Czas trwania wypowiedzi ulega kilkukrotnemu wydłużeniu.



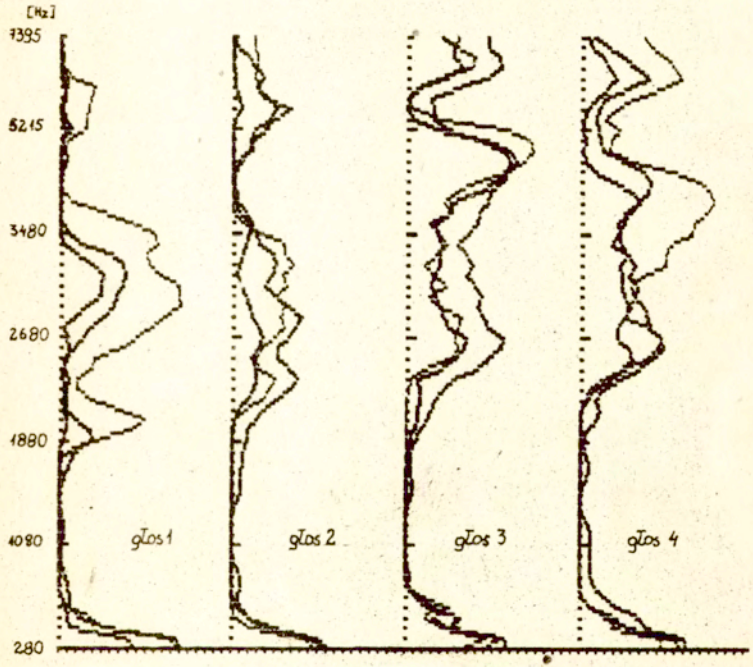
Rys. 1 Samogłoski polskie na czworoboku samogłoskowym



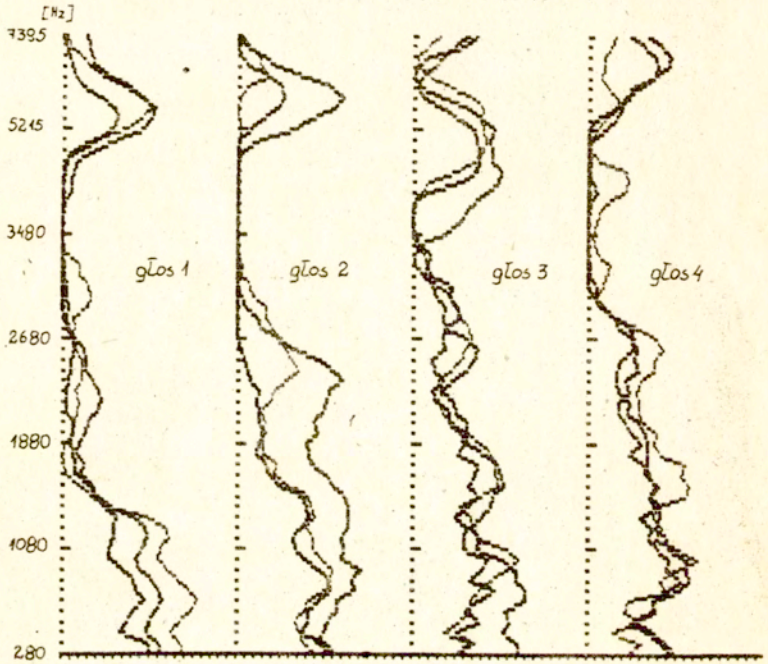
Rys. 2 Średnie częstotliwości formantów F_1 i F_2 dla poszczególnych samogłosek [Jassem 1984]



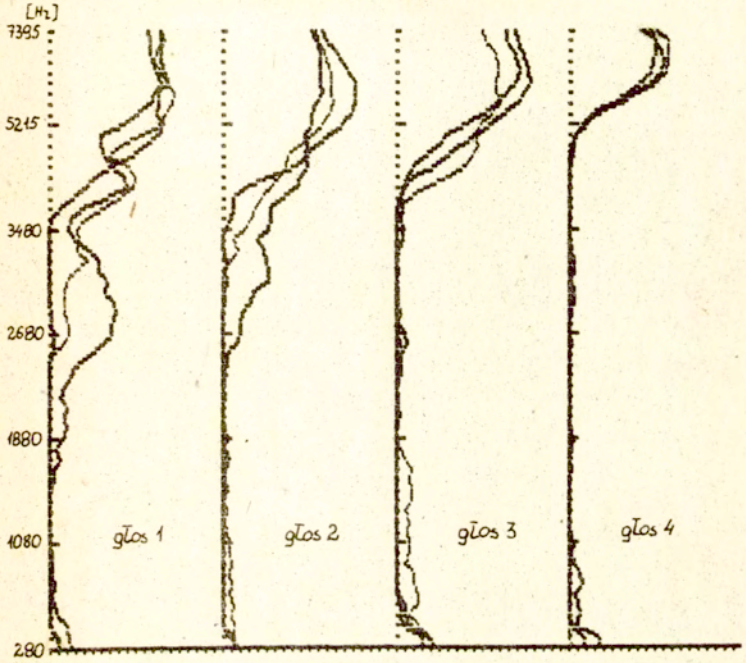
Rys.3 Graf konstrukcji fonokodu



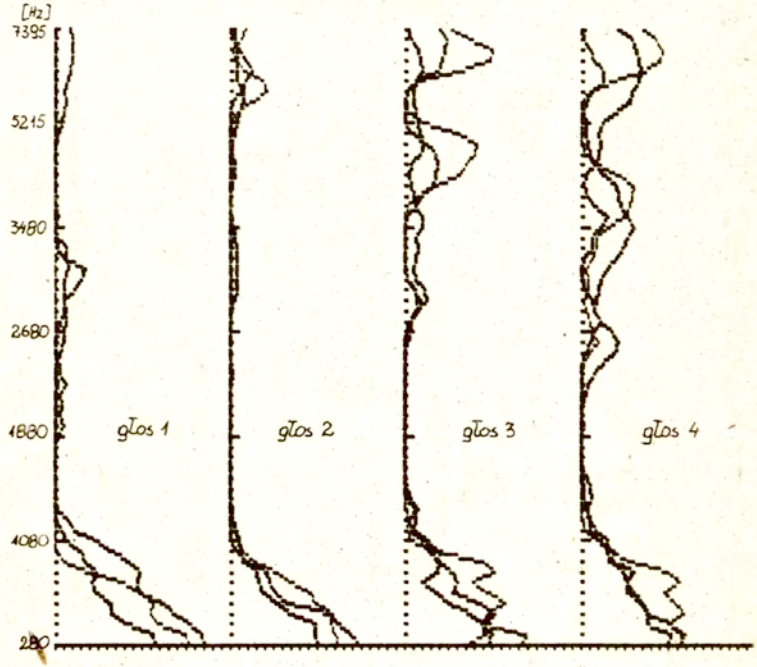
Rys.4a Średnie widma segmentów - głoska i



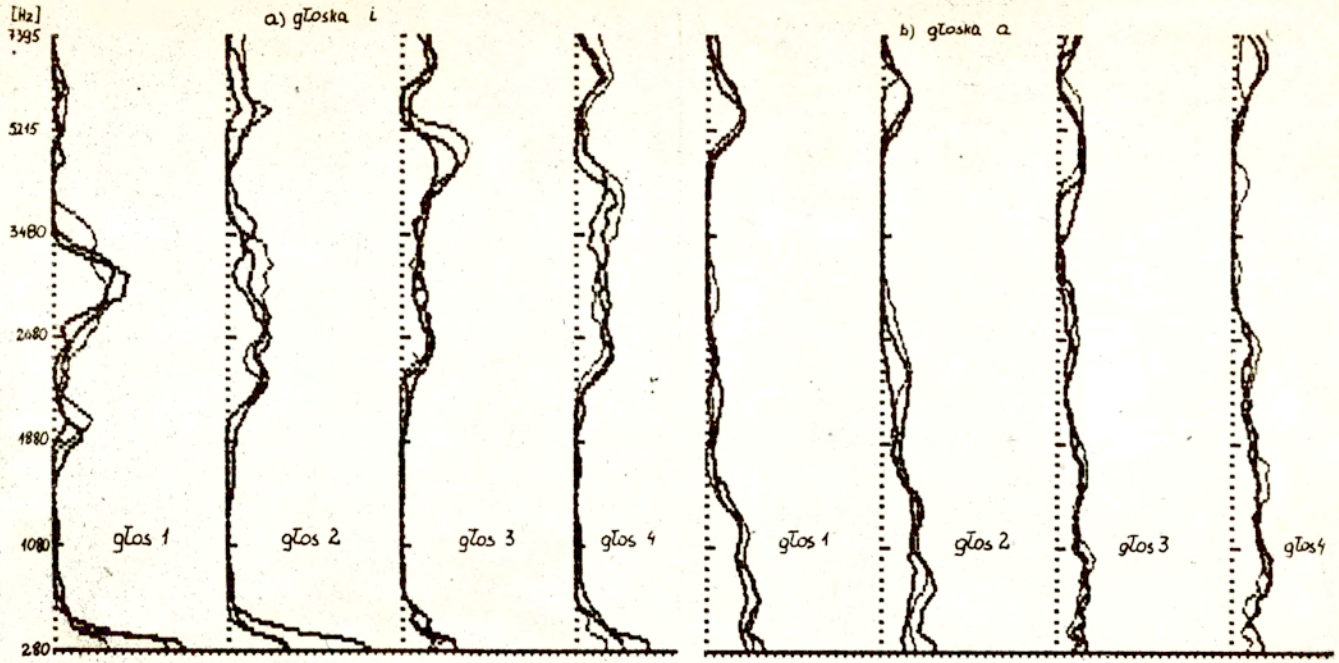
Rys.4b Średnie widma segmentów - głoska a



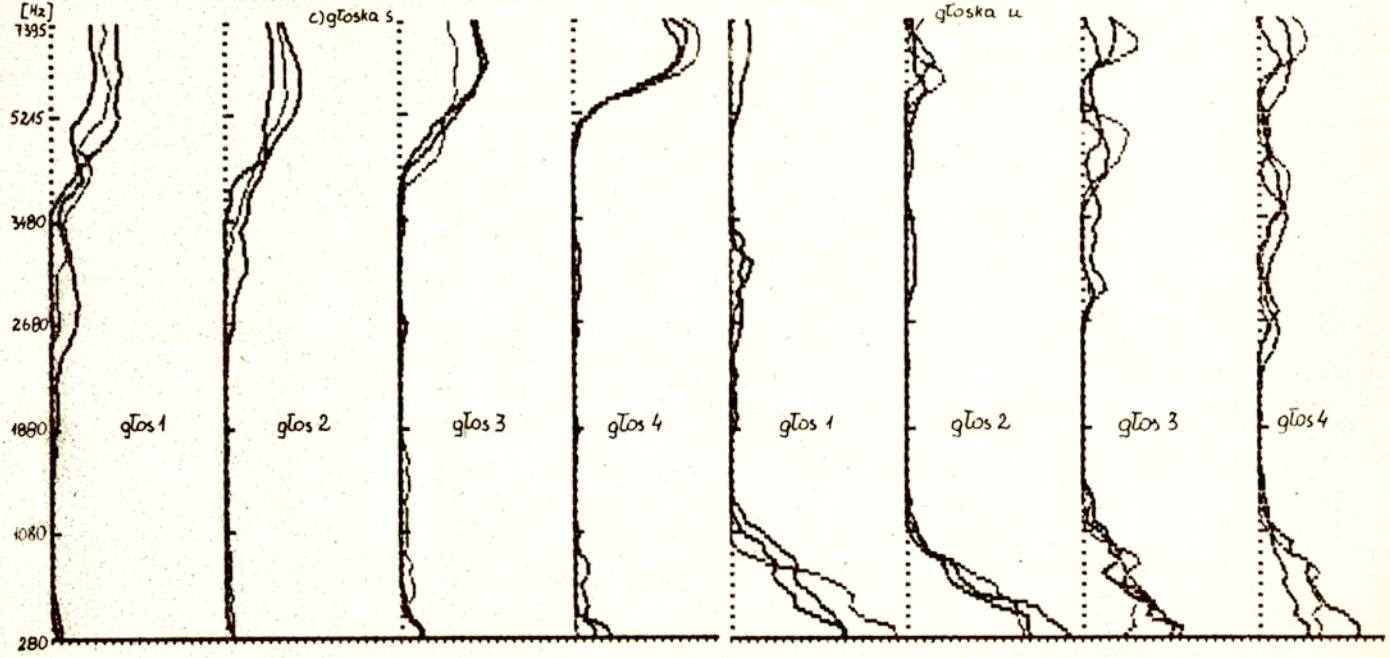
Rys. 4c Średnie widma segmentów - głoska a



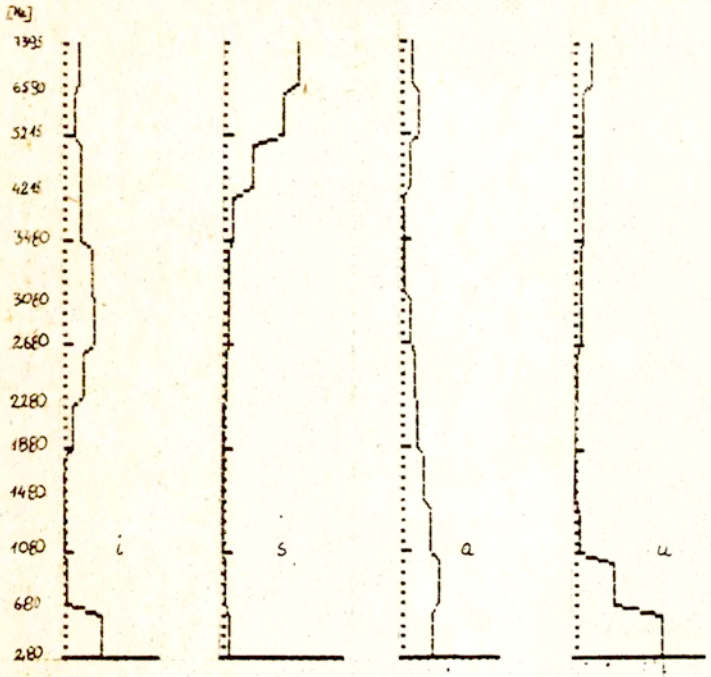
Rys. 4d Średnie widma segmentów - głoska u



Rys 5 a,b Średnie, unormowane widma głosek



Rys. 5c,d Średnie, unormowane widma głosek



Rys.6 Unormowane, zredukowane reprezentacje głosek

Tablica 1.

Liczba fonemów	1	2	3	4	5	6	7	8	9	10
Liczebność fonokodu	4	5	24	50	169	420	1264	3365	9684	26510

Tablica 2. Wyniki rozpoznawania fonokodu.

Lp	Fonokod	Głosy															
		LR		HK		PD		AM		KG		IW		RB		BI	
		ż		mw		mw		m		m		zw		ż		ż	
1	I	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
2	A	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
3	U	+	+	+	+	+	+	+	+	IU	IU	+	+	+	+	+	+
4	S	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
5	IS	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
6	AS	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
7	US	+	+	+	+	UIS	+	+	+	+	+	+	+	+	+	+	+
8	SA	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
9	SU	+	+	SIU	+	+	+	+	+	+	+	SIU	+	+	+	+	+
10	IPI	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
11	ISU	+	+	+	+	ISIU	+	+	+	+	+	+	+	+	+	ISAU	+
12	ASA	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
13	SPS	+	+	+	+	+	+	SPIS	+	+	+	+	+	+	+	+	+
14	APU	+	+	+	+	+	+	+	+	+	+	+	+	APSU	+	+	+
15	UPI	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
16	ISA	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
17	SAS	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
18	SPU	+	+	+	+	+	+	SPSU	+	+	+	+	+	+	+	+	+

ż - głos żeński, m - głos męski, w - głos wzorcowy,
\+ - rozpoznanie poprawne

BIBLIOGRAFIA

1 DOMAGAŁA, P., Automatyzacja procesu segmentacji sygnału mowy w układzie analogowo-cyfrowym, Prace IPPT 5/1984, Warszawa, 1984.

2 DREYFUS-GRAF, J., A., Coded speech Phonocodes and recognition machines, Proc. 8th I.C.A., London, 1974.

3 JASSEM, W., Podstawy fonetyki akustycznej, PWN, Warszawa, 1973.

4 JASSEM, W., KRZYSKO, M., STOLARSKI, P., Częstotliwości formantowe samogłosek jako cechy fonematyczne i osobnicze w świetle statystycznej analizy dyskryminacyjnej, Prace IPPT 27/84, Warszawa, 1984.

5 KOSTER, J., P., DREYFUS-GRAF, J., A., Phonocodes und die perception Konstruierter Sprachen, Hamburger Phonetische Beitrage, Miszellen III, Helmut Verlag, Band 17, Hamburg, 1976.

6 JASSEM, W., Wstępne założenia akustycznej teorii fonemu, Materiały XXXII Otwartego Seminarium z Akustyki, OSA 1985, Kraków, 1985, 61-64.

	PONOKOD				
	I	A	S	U	
1	IS	AS	SA	SU	US
3	IPI	IPA	IPS	IPU	ISA
	ISU	API	APA	APS	APU
	ASA	ASU	SPI	SPA	SPS
	SPU	SAS	SUS	UPI	UPA
	UPS	UPU	USA	USU	
4	IPIS	IPAS	IPSA	IPSU	IPUS
	ISPI	ISPA	ISPS	ISPU	ISAS
	ISUS	APIS	APAS	APSA	APSU
	APUS	ASPI	ASPA	ASPS	ASPU
	ASAS	ASUS	SPIS	SPAS	SPSA
	SPSU	SPUS	SAPI	SAPA	SAPS
	SAPU	SASA	SASU	SUPI	SUPA
	SUPS	SUPU	SUSA	SUSU	UPIS
	UPAS	UPSA	UPSU	UPUS	USPI
	USPA	USPS	USPU	USAS	USUS
5	IPIPI	IPIPA	IPIPS	IPIPU	IPISA
	IPISU	IPAPI	IPAPA	IPAPS	IPAPU
	IPASA	IPASU	IPSPI	IPSPA	IPSPS
	IPSPU	IPSAS	IPSUS	IPUPI	IPUPA
	IPUPS	IPUPU	IPUSA	IPUSU	ISPIS
	ISPAS	ISPSA	ISPSU	ISPUS	ISAPI
	ISAPA	ISAPS	ISAPU	ISASA	ISASU
	ISUPI	ISUPA	ISUPS	ISUPU	ISUSA
	ISUSU	APIPI	APIPA	APIPS	APIPU
	APISA	APISU	APAPI	APAPA	APAPS
	APAPU	APASA	APASU	APSPI	APSPA
	APSPS	APSPU	APSAS	APSUS	APUPI
	APUPA	APUPS	APUPU	APUSA	APUSU
	ASPIS	ASPAS	ASPSA	ASPSU	ASPUS
	ASAPI	ASAPA	ASAPS	ASAPU	ASASA
	ASASU	ASUPI	ASUPA	ASUPS	ASUPU
	ASUSA	ASUSU	SPIPI	SPIPA	SPIPS
	SPIPU	SPISA	SPISU	SPAPI	SPAPA
	SPAPS	SPAPU	SPASA	SPASU	SPSPI
	SPSPA	SPSPS	SPSPU	SPSAS	SPSUS
	SPUPI	SPUPA	SPUPS	SPUPU	SPUSA
	SPUSU	SAPIS	SAPAS	SAPSA	SAPSU
	SAPUS	SASPI	SASPA	SASPS	SASPU
	SASAS	SASUS	SUPIS	SUPAS	SUPSA
	SUPSU	SUPUS	SUSPI	SUSPA	SUSPS
	SUSPU	SUSAS	SUSUS	UPIPI	UPIPA
	UPIPS	UPIPU	UPISA	UPISU	UPAPI
	UPAPA	UPAPS	UPAPU	UPASA	UPASU
	UPSPI	UPSPA	UPSPS	UPSPU	UPSAS
	UPSUS	UPUPI	UPUPA	UPUPS	UPUPU
	UPUSA	UPUSU	USPIS	USPAS	USPSA
	USPSU	USPUS	USAPI	USAPA	USAPS
	USAPU	USASA	USASU	USUPI	USUPA
	USUPS	USUPU	USUSA	USUSU	